W770–W773 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki432

Current Comparative Table (CCT) automates customized searches of dynamic biological databases

Benjamin R. Landsteiner, Michael R. Olson and Robert Rutherford*

Department of Biology, St Olaf College, 1520 St Olaf Avenue, Northfield, MN 55057, USA

Received February 14, 2005; Revised and Accepted March 24, 2005

ABSTRACT

The Current Comparative Table (CCT) software program enables working biologists to automate customized bioinformatics searches, typically of remote sequence or HMM (hidden Markov model) databases. CCT currently supports BLAST, hmmpfam and other programs useful for gene and ortholog identification. The software is web based, has a BioPerl core and can be used remotely via a browser or locally on Mac OS X or Linux machines. CCT is particularly useful to scientists who study large sets of molecules in today's evolving information landscape because it color-codes all result files by age and highlights even tiny changes in sequence or annotation. By empowering non-bioinformaticians to automate custom searches and examine current results in context at a glance, CCT allows a remote database submission in the evening to influence the next morning's bench experiment. A demonstration of CCT is available at http://orb.public.stolaf.edu/CCTdemo and the open source software is freely available from http://sourceforge.net/projects/orb-cct.

INTRODUCTION

A constant flood of new genomic information has brought a new age of discovery to biology. Unfortunately, this deluge of new data is rarely fully utilized, in part because bench scientists find it increasingly challenging to maintain a current, integrated picture of the latest data. This problem is particularly pronounced for scientists who study large numbers of potentially interesting molecules, a common result of microarray-based or proteome-based experiments. Regularly updating such data by hand can be extremely burdensome and is rarely done. As a result, many scientists work unaware of newly discovered annotation, homologs, clones or protein domains that could further their research.

Current Comparative Table, or CCT, is a web-based application designed to solve this problem by displaying the most up-to-date results of customized genomic searches in a convenient table. It can be easily configured to automatically download new versions of databases and to run any number of bioinformatics searches on the new data. CCT then organizes the results in a table containing hyperlinks color-coded by result age, making it simple to pick out recently changed results. This allows scientists to continuously harvest potentially useful data about any set of sequences of importance to their research.

CCT joins a number of other excellent servers whose aim is to manage data overload (1,2), but CCT has unique strengths. The most widely used of these other services, PubCrawler, searches for text and literature matches in PubMed, GenBank or both but does not perform actual sequence searching (3). Servers that do allow sequence searches (e.g. http://www.expasy.org/swiss-shop and http://myhits.isb-sib.ch) allow the searching of prescribed subsets of public databases and/or a single private database of each data type that is uploaded manually (2). Like MyHits and Swiss-Shop, CCT performs sequence-based searches, but it uniquely offers a local installation, allowing independence from remote servers. It is also simple to customize because it is open source software. Other novel features of CCT include automated highlighting of even small changes in data files, a simple interface for scientists interested in multiple sequences and the ability to monitor any number of databases to which the user has access.

MATERIALS AND METHODS

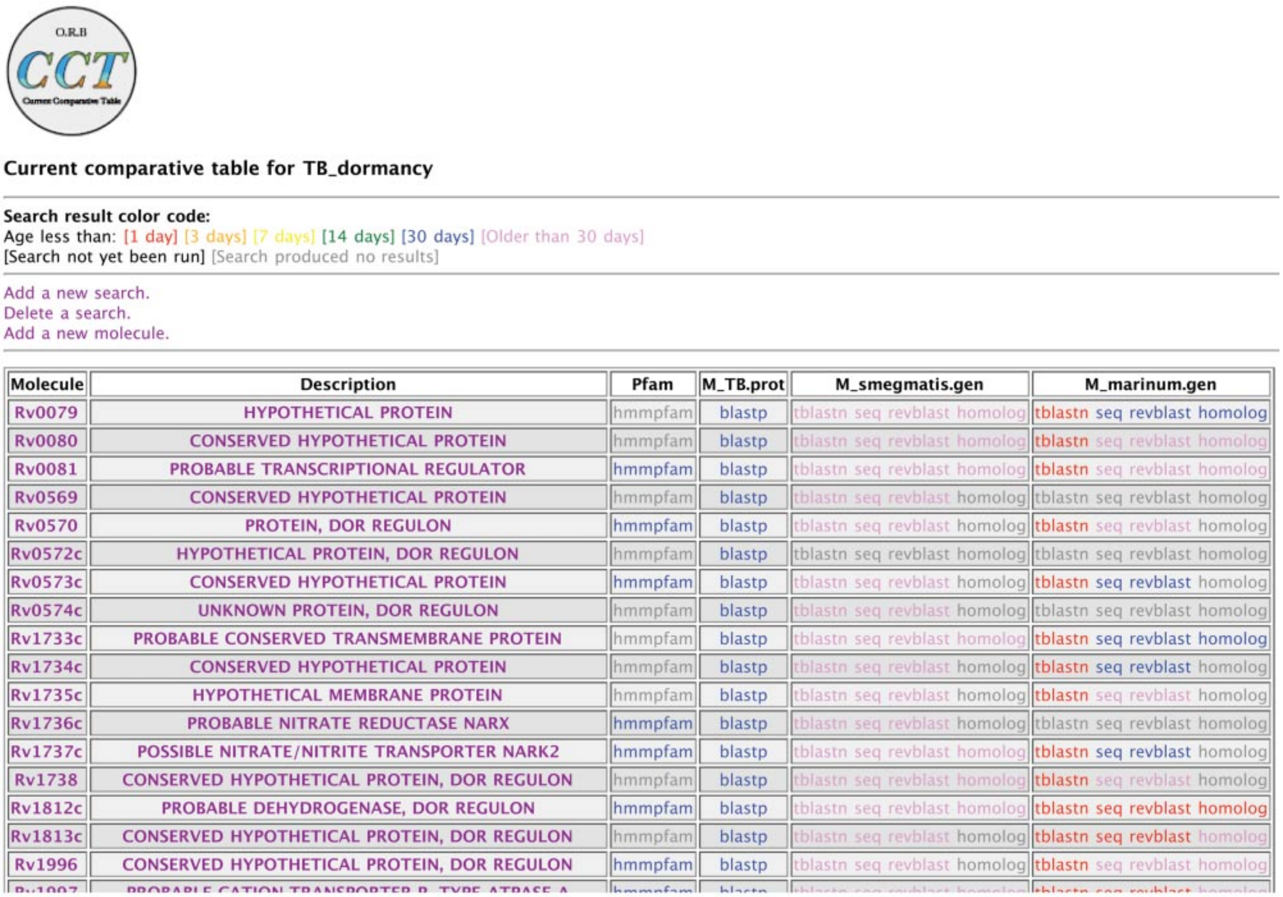
CCT was developed on a Gateway E-6100 series computer running RedHat 9 Linux. The computer has a 3 GHz processor, 200 GB of hard drive space and 2 GB of RAM. Mac OS X compatibility was tested on a Dual 2 GHz G5 Tower with 1.5 GB of RAM, a 160 GB hard drive and Mac OS X 10.3.4. CCT is implemented in Perl and makes heavy use of the BioPerl toolkit (4). CCT is freely available and open source. A demonstration of CCT and an installation guide are available at http://orb.public.stolaf.edu/CCTdemo, and the software is freely available from http://sourceforge.net/projects/orb-cct.

*To whom correspondence should be addressed. Tel: +1 507 646 3804; Fax: +1 507 646 3968; Email: robruth@stolaf.edu

USAGE

A scientist typically begins using CCT by adding three types of data through a web interface: (i) a file containing sequences of interest, (ii) database location(s) to monitor for updates and (iii) searches to perform. Although CCT can be run at will, it is typically run automatically. In this mode, CCT periodically (e.g. once per day, such as at 1 a.m.) checks user-selected databases and downloads updated versions as they become available. CCT then runs user-selected searches on these data and builds a table with one row for each sequence and one column for each database searched (Figure 1). Each cell contains links to the results of each search, color-coded by the length of time since a change in data has affected the search results. In addition, when a result is updated, the new result is compared with the previous one, and differences are highlighted in the new data file (Figure 2). Taken together, link coloring and difference highlighting allow numerous search results to be quickly scanned and evaluated for novelty (or for stability over time). This feature is valuable for scientists engaged in ongoing projects as well as those deciding when to commit limited laboratory resources to the characterization of a set of interesting but preliminary sequences.

CCT can be run locally or remotely via the Internet and comes bundled with wrappers for six different bioinformatics search programs. hmmpfam is a wrapper for the hmmpfam tool and is useful for automating searches of sequences against ever changing protein domain databases such as Pfam (5). tblastn and blastp are the software's wrappers for NCBI's BLAST searches (6). These programs use the local tool blastall to run searches against a sequence database. The seq program is specific to CCT and is useful for isolating regions of the genome, e.g. for finding genes and open reading frames. seq takes tblastn or blastp output, captures the sequences for the target BLAST high-scoring pairs and extends them out to a user-specified delimiter. For example, if a stop codon were selected, seq would capture the sequence of the flanking open reading frame. revblast uses seq output as a query to BLAST search another database, permitting reciprocal blasting, a common method for finding ortholog pairs (7). Finally, the homolog program takes revblast output and uses Clustal-W (8) to generate a pairwise alignment if sequence pairs meet



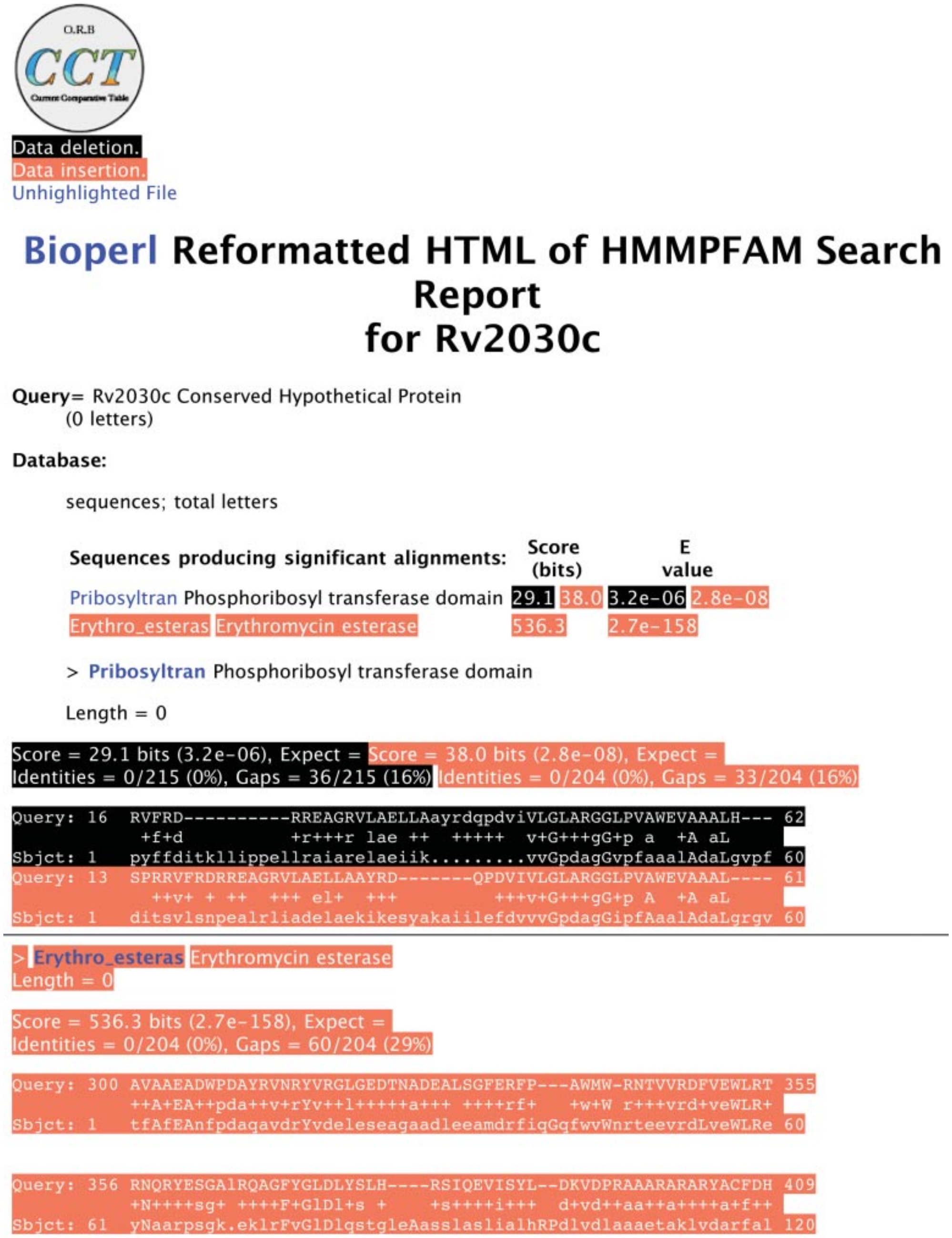
Current comparative table for TB_dormancy

Search result color code:
Age less than: [1 day] [3 days] [7 days] [14 days] [30 days] [Older than 30 days]
[Search not yet been run] [Search produced no results]

Add a new search.
Delete a search.
Add a new molecule.

Molecule	Description	Pfam	M_TB.prot	M_smegmatis.gen	M_marinum.gen
Rv0079	HYPOTHETICAL PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0080	CONSERVED HYPOTHETICAL PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0081	PROBABLE TRANSCRIPTIONAL REGULATOR	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0569	CONSERVED HYPOTHETICAL PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0570	PROTEIN, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0572c	HYPOTHETICAL PROTEIN, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0573c	CONSERVED HYPOTHETICAL PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv0574c	UNKNOWN PROTEIN, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1733c	PROBABLE CONSERVED TRANSMEMBRANE PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1734c	CONSERVED HYPOTHETICAL PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1735c	HYPOTHETICAL MEMBRANE PROTEIN	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1736c	PROBABLE NITRATE REDUCTASE NARX	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1737c	POSSIBLE NITRATE/NITRITE TRANSPORTER NARK2	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1738	CONSERVED HYPOTHETICAL PROTEIN, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1812c	PROBABLE DEHYDROGENASE, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1813c	CONSERVED HYPOTHETICAL PROTEIN, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1996	CONSERVED HYPOTHETICAL PROTEIN, DOR REGULON	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog
Rv1997	PROBABLE CATION TRANSPORTER P-TYPE ATPASE A	hmmpfam	blastp	tblastn seq revblast homolog	tblastn seq revblast homolog

Figure 1. A small part of a sample current comparative table. The molecule and description columns contain information supplied by the user; the other columns contain data generated by automated searches against user-specified databases. In these database columns, the cells contain links to the result of each search. Links are color-coded to reflect the length of time since the search result last changed. To generate the table shown, CCT was given a set of 48 genes in the *Mycobacterium tuberculosis* (MTB) Dormancy regulon (10) and directed to monitor four remote databases (Pfam, the MTB proteome and two incomplete genomes, *M.smegmatis* and *M.marinum*). The pattern of red indicates that a new release of the *M.marinum* genome was downloaded in the past 24 h (generating many new red tblastn links), and that a new sequence (seq program) of an ortholog (homolog program) was found for the TB protein Rv1812c. The predicted ortholog for Rv0079, in contrast, did not change as a result of the update and the corresponding links remain blue. More detail concerning any result can be viewed by clicking on the corresponding link(s).



Data deletion.
Data insertion.
Unhighlighted File

Bioperl Reformatted HTML of HMMPFAM Search Report for Rv2030c

Query= Rv2030c Conserved Hypothetical Protein
(0 letters)

Database:

sequences; total letters

```
Sequences producing significant alignments:     Score       E
                                                (bits)     value
Pribosyltran Phosphoribosyl transferase domain  29.1 38.0  3.2e-06 2.8e-08
Erythro_esteras Erythromycin esterase           536.3      2.7e-158
```

> **Pribosyltran** Phosphoribosyl transferase domain

Length = 0

Score = 29.1 bits (3.2e-06), Expect = Score = 38.0 bits (2.8e-08), Expect =
Identities = 0/215 (0%), Gaps = 36/215 (16%) Identities = 0/204 (0%), Gaps = 33/204 (16%)

```
Query: 16  RVFRD----------RREAGRVLAELLAayrdqpdviVLGLARGGLPVAWEVAAALH--- 62
             +f+d          +r+++r lae ++  +++++  v+G+++gG+p a  +A aL
Sbjct: 1   pyffditkllippellraiarelaeiik.........vvGpdagGvpfaaalAdaLgvpf 60
Query: 13  SPRRVFRDRREAGRVLAELLAAYRD-------QPDVIVLGLARGGLPVAWEVAAAL---- 61
              ++v+ + ++  +++ el+  +++         +++v+G+++gG+p A  +A aL
Sbjct: 1   ditsvlsnpealrliadelaekikesyakaiilefdvvvGpdagGipfAaalAdaLgrgv 60
```

> Erythro_esteras Erythromycin esterase
Length = 0

Score = 536.3 bits (2.7e-158), Expect =
Identities = 0/204 (0%), Gaps = 60/204 (29%)

```
Query: 300 AVAAEADWPDAYRVNRYVRGLGEDTNADEALSGFERFP---AWMW-RNTVVRDFVEWLRT 355
            ++A+EA++pda++v+rYv++l+++++a+++ ++++rf+   +w+W r+++vrd+veWLR+
Sbjct: 1   tfAfEAnfpdaqavdrYvdeleseagaadleeamdrfiqGqfwvWnrteevrdLveWLRe 60


Query: 356 RNQRYESGAlRQAGFYGLDLYSLH----RSIQEVISYL--DKVDPRAAARARARYACFDH 409
            +N++++sg+ ++++F+GlDl+s +    +s++++i+++  d+vd++aa++a++++a+f++
Sbjct: 61  yNaarpsgk.eklrFvGlDlqstgleAasslaslialhRPdlvdlaaaetaklvdarfal 120
```

Figure 2. Example of a result file revised to reflect a new release of the Pfam database. Red highlighting indicates new data; black highlighting shows removed data. The regions above and below the horizontal line show different parts of the same result file. This hmmpfam search result shows that the 'Conserved Hypothetical Protein' Rv2030c matches a new domain in Pfam, specifically the 'erythromycin esterase' domain. CCT's highlighting also shows a second finding: the model for the 'phosphoribosyl' domain has been changed subtly (note the changed amino acids in the subject line). Taken together with the change in database size for the new release, the *E*-value for this search has changed somewhat. To view the newest results only (without highlighting to show changes) a user can click on the 'Unhighlighted File' link.

user-specified parameters. When used together, these programs are an effective tool for finding orthologs and can be customized in a way that sets CCT apart from other comparative genomics tools (9).

The Mac OS X version of CCT can be installed in ~60 s in a few simple steps from a double clickable install package. It is fully self-contained and includes code for BioPerl, Clustal-W, blastall and hmmpfam. The Linux version does not include this code in the expectation that Linux users may want to integrate CCT into existing bioinformatics resources on their servers.

CCT's user manual is part of every installation and can also be found at the software's demo site. The user manual contains an installation guide, a beginner's guide, the addresses of sample databases, screenshots and other useful information. In addition, CCT is installed with a link to extensive code documentation to make its customization as easy as possible for users with any level of programming experience.

Programmers can construct new program modules to interact seamlessly with CCT using an included template file.

The design of CCT falls into two main parts: the web interface and the script runCCT.pl. The web interface uses the Perl CGI to interact with the browser. Users can add and delete tables, searches and databases, and can view their data by browsing a CCT web page. runCCT.pl controls most of CCT's daily work, such as downloading databases, running searches and updating tables. This script can also be called manually from the command line and can be manipulated to perform only certain steps of its process or run only specified searches.

CCT is freely available to all and it will continue to be developed (http://sourceforge.net/projects/orb-cct). Users who find it especially valuable may cite this publication.

CCT can be a very useful tool for scientists who study large sets of genes in today's evolving genomic landscape. By empowering non-bioinformaticians to automate custom searches and examine current results at a glance, CCT allows a remote database submission in the evening to influence the next morning's bench experiment.

ACKNOWLEDGEMENTS

The authors thank the BioPerl consortium and SourceForge for support, Bob Bried for system administration, E. Savage and E. Zweifel and many others for testing, The Institute for Genomic Research and the Sanger Center for access to incomplete genomes and our referees and colleagues for constructive feedback. M.R.O. and B.R.L. were supported by a Howard Hughes Medical Institute Undergraduate Science Education grant. Funding to pay the Open Access publication charges for this article was provided by St Olaf College.

Conflict of interest statement. None declared.

REFERENCES

1. Hokamp,K. and Wolfe,K. (1999) What's new in the library? What's new in GenBank? Let PubCrawler tell you. *Trends Genet*., **15**, 471–472.
2. Pagni,M., Ioannidis,V., Cerutti,L., Zahn-Zabal,M., Jongeneel,C.V. and Falquet,L. (2004) MyHits: a new interactive resource for protein annotation and domain identification. *Nucleic Acids Res*., **32**, W332–W335.
3. Hokamp,K. and Wolfe,K.H. (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. *Nucleic Acids Res*., **32**, W16–W19.
4. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
5. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol*., **215**, 403–410.
7. Hirsh,A.E. and Fraser,H.B. (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
9. Frazer,K.A., Elnitski,L., Church,D.M., Dubchak,I. and Hardison,R.C. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res*., **13**, 1–12.
10. Voskuil,M.I., Schnappinger,D., Visconti,K.C., Harrell,M.I., Dolganov,G.M., Sherman,D.R. and Schoolnik,G.K. (2003) Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program. *J. Exp. Med*., **198**, 705–713.